Variable-Rate Source Coding Theorems for Stationary Nonergodic Sources

M. Effros, *Member, IEEE,* P. A. Chou, *Member, IEEE,* and R. M. Gray, *Fellow, IEEE*

***Abstract*—For a stationary ergodic source, the source coding theorem and its converse imply that the optimal performance theoretically achievable by a fixed-rate or variable-rate block quantizer is equal to the distortion-rate function, which is defined as the infimum of an expected distortion subject to a mutual information constraint. For a stationary nonergodic source, however, the distortion-rate function cannot in general be achieved arbitrarily closely by a fixed-rate block code. We show, though, that for any stationary nonergodic source with a Polish alphabet, the distortion-rate function can be achieved arbitrarily closely by a variable-rate block code. We also show that the distortion-rate function of a stationary nonergodic source has a decomposition as the average of the distortion-rate functions of the source's stationary ergodic components, where the average is taken over points on the component distortion-rate functions having the same slope. These results extend previously known results for finite alphabets.**

***Index Terms*—Source coding theorems, stationary nonergodic sources, distortion-rate function.**

I. Introduction

In [1], Shields, Neuhoff, Davisson, and Ledrappier show that for any stationary nonergodic source with a finite alphabet, the distortion-rate function $D(R)$ equals the infimum of the average of the distortion-rate functions of the source's stationary ergodic components, where the average is taken over points on the component distortion-rate functions whose rates, on average, are at most R. Leon-Garcia, Davisson, and Neuhoff [2] later prove the achievability of this bound by variable-rate block codes.

In this work we extend these variable-rate quantization results from finite alphabets to complete separable metric spaces, or Polish alphabets, of which finite alphabets are a special case. As in the previous works, we employ a variable-rate and variable-distortion approach. However, we simplify the approach using a Lagrangian formulation. Our methods of proof rely heavily on theorems from [3] and [4], which are stated here but not proved.

Manuscript received July 6, 1993; revised March 31, 1994. This material is based upon work partially supported by the National Science Foundation under an NSF Graduate Fellowship, by a grant from The Center for Telecommunications at Stanford, and by an AT & T Ph.D. Scholarship. This paper was presented in part at the 1994 IEEE International Symposium on Information Theory, Trondheim, Norway.

M. Effros was formerly with the Information Systems Laboratory, Stanford, CA 94305-4055. She is currently with the Department of Electrical Engineering (116-81), California Institute of Technology, Pasadena, CA 91125.

P. A. Chou is with the Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304.

R. M. Gray is with the Information Systems Laboratory, Stanford, CA 94305-4055.

IEEE Log Number 9406004.

II. Results

Let $(A^\infty, \mathscr{B}^\infty, \mu, T)$ be a stationary dynamical system with Polish alphabet A. That is, let A be a complete separable metric space, let $\mathscr{B}$ be the Borel σ-algebra generated by the open sets of A, let A^∞ be the set of one-sided sequences $x = (x_1, x_2, \cdots)$ from A, let $\mathscr{B}^\infty$ be the σ-algebra of subsets of A^∞ generated by finite-dimensional rectangles with components in $\mathscr{B}$, let T be the left shift operator on A^∞, and let μ be a measure on the measurable space $(A^\infty, \mathscr{B}^\infty)$, stationary with respect to T.

Now let $\rho(x_1, y_1) < \infty$ be a real-valued nonnegative distortion measure for $x_1 \in A$, $y_1 \in \hat{A}$, where $\hat{A}$ is an abstract reproduction alphabet. Assume that $\rho(x_1, y_1)$ is continuous in x_1 for each $y_1 \in \hat{A}$ and that there exists a reference letter y_1^* such that $E_\mu \rho(X_1, y_1^*) < \infty$. Define $\rho(x^N, y^N) = \sum_{i=1}^N \rho(x_i, y_i)$.

Finally, let Q be a variable-rate block quantizer with blocklength N. That is, let Q be a map from A^N onto some finite or countable set of codewords $\{y^N\} \subseteq A^N$ composing a codebook $\mathscr{C} = \{(y^N, |y^N|)\}$ in which each codeword y^N has an associated variable-length binary description, with length denoted $|y^N|$. The description lengths must satisfy the Kraft inequality $\sum_{y^N \in \mathscr{C}} 2^{-|y^N|} \leq 1$.

The optimal performance theoretically achievable by any variable-rate block quantizer is the operational distortion-rate function

$$\delta^{\mathrm{vr}}(R, \mu) = \inf_N \delta_N^{\mathrm{vr}}(R, \mu), \tag{1}$$

where $\delta_N^{\mathrm{vr}}(R, \mu)$ is the Nth-order operational distortion-rate function

$$\delta_N^{\mathrm{vr}}(R, \mu) = \inf_Q \left\{ \frac{1}{N} E_\mu \rho(X^N, Q(X^N)) : \frac{1}{N} E_\mu |Q(X^N)| \leq R \right\}. \tag{2}$$

Here, the infimum is taken over all variable-rate block quantizers Q with blocklength N. We pointedly distinguish this from the optimal performance theoretically achievable by fixed-rate block quantizers, $\delta^{\mathrm{fr}}(R, \mu) = \inf_N \delta_N^{\mathrm{fr}}(R, \mu)$, in which $\delta_N^{\mathrm{fr}}(R, \mu)$ is defined as in (2) but with the infimum taken over all fixed-rate block quantizers with blocklength N.

The Shannon distortion-rate function is defined similarly, as

$$D(R, \mu) = \inf_N D_N(R, \mu), \tag{3}$$

where $D_N(R, \mu)$ is the Nth-order distortion-rate function

$$D_N(R, \mu) = \inf_\nu \left\{ \frac{1}{N} E_{\mu\nu} \rho(X^N, Y^N): \frac{1}{N} I_{\mu\nu}(X^N; Y^N) \le R \right\}. \tag{4}$$

Here, ν is a conditional probability or test channel from A^N to $\hat{A}^N$ defining, with μ, a joint probability or hookup $\mu\nu$ on X^N and Y^N, and I is the mutual information.

It is well known that both $D(R, \mu)$ and $\delta^{\text{fr}}(R, \mu)$ are convex in R [4]–[7]; $\delta^{\text{vr}}(R, \mu)$ is also convex in R, for the same reason as $\delta^{\text{fr}}(R, \mu)$ (a time-sharing argument). Hence $\delta^{\text{vr}}(R, \mu)$ and $D(R, \mu)$ can be characterized by their support functionals [8, p. 135]

$$l(\lambda, \mu) = \inf_R [\delta^{\text{vr}}(R, \mu) + \lambda R] \tag{5}$$

and

$$L(\lambda, \mu) = \inf_R [D(R, \mu) + \lambda R]. \tag{6}$$

$L(\lambda, \mu)$ can be interpreted in graphical terms as the y-intercept of the line supporting the graph of $D(R, \mu)$ at slope $-\lambda$, as in Fig. 1; $l(\lambda, \mu)$ can be interpreted similarly in terms of $\delta^{\text{vr}}(R, \mu)$. These functionals can also be interpreted as Lagrangians with Lagrange multiplier λ. We shall call them the weighted operational distortion-rate function and the weighted Shannon distortion-rate function, respectively.

The source coding theorem and its converse [6, Theorems 7.2.4, 7.2.5] imply that when μ is ergodic, $\delta^{\text{vr}}(R, \mu) = \delta^{\text{fr}}(R, \mu) = D(R, \mu)$ for all $R \ge 0$ (and hence $l(\lambda, \mu) = L(\lambda, \mu)$ for all $\lambda \ge 0$). When μ is nonergodic, let $\{\mu_x: x \in A^\infty\}$ denote the ergodic decomposition of μ. The ergodic decomposition exists since, if A is Polish, then $(A, \mathcal{B})$ is standard [3, Theorem 3.3.1], and hence $(A^\infty, \mathcal{B}^\infty)$ is standard [3, Lemma 2.4.1]; standard measurable spaces admit the ergodic decomposition [3, Theorem 7.4.1]. The main results of this paper are that under the conditions given above (namely, A is Polish, $\rho(x_1, y_1)$ is continuous in x_1 for each y_1, and there exists a reference letter y_1^* such that $E_\mu \rho(X_1, y_1^*) < \infty$), the following hold.

Theorem 1: $l(\lambda, \mu) = \int l(\lambda, \mu_x)\, d\mu(x)$, $\forall \lambda \ge 0$.

Theorem 2: $L(\lambda, \mu) = \int L(\lambda, \mu_x)\, d\mu(x)$, $\forall \lambda \ge 0$.

Theorem 3: $l(\lambda, \mu) = L(\lambda, \mu)$, $\forall \lambda \ge 0$, and hence $\delta^{\text{vr}}(R, \mu) = D(R, \mu)$, $\forall R \ge 0$.

Theorem 3 results from Theorems 1 and 2 since $l(\lambda, \mu_x) = L(\lambda, \mu_x)$ for each stationary ergodic source μ_x. The implications of these results are discussed in Section III, followed by the proofs, given in a series of lemmas, in Section IV. Many of the results of this paper arise from results in [3] and [4]. Those theorems are labeled and included in Section IV as well.

III. Discussion

Theorems 1 and 2 are known as ergodic decompositions of the functionals $l(\lambda, \mu)$ and $L(\lambda, \mu)$. The theorems say that for any λ, the weighted operational distortion-rate function $l(\lambda, \mu)$ and the weighted Shannon distortion-rate function $L(\lambda, \mu)$ of a stationary nonergodic source μ can each be found as the expected value of the corresponding functionals of the ergodic subsources $\{\mu_x\}$. These theorems can also be expressed directly in terms of the operational distortion-rate function $\delta^{\text{vr}}(R, \mu)$ and the Shannon distortion-rate function $D(R, \mu)$. Theorem 2, for example, can be expressed

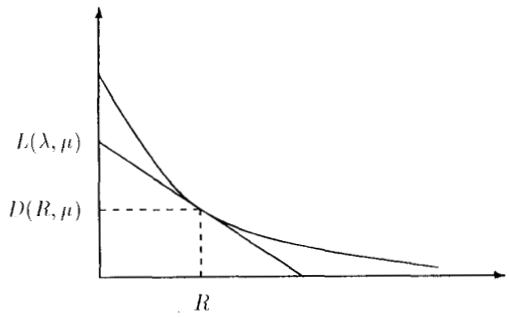

Fig. 1. Graphical interpretation of $L(\lambda, \mu)$.

$$\begin{aligned} L(\lambda, \mu) &= \int \inf_R [D(R, \mu_x) + \lambda R]\, d\mu(x) \\ &= \inf_{\{R_x\}} \int [D(R_x, \mu_x) + \lambda R_x]\, d\mu(x) \\ &= \inf_{\{R_x\}} \left[\int D(R_x, \mu_x)\, d\mu(x) + \lambda \int R_x\, d\mu(x) \right], \end{aligned}$$

where a theorem of Chung [9] allows the integral and the infimum to be exchanged (and furthermore allows the infimum over $\{R_x\}$ to be restricted to measurable functions of x if $D(R, \mu_x)$ is a measurable function of R and x). Thus $L(\lambda, \mu)$ can be interpreted as a Lagrangian for the minimization of $\int D(R_x, \mu_x)\, d\mu(x)$ subject to a constraint on $\int R_x\, d\mu(x)$. Hence by the Lagrange duality theorem [8, p. 224],

$$\max_{\lambda \ge 0} [L(\lambda, \mu) - \lambda R_0] = \inf_{\{R_x\}} \left\{ \int D(R_x, \mu_x)\, d\mu(x): \int R_x\, d\mu(x) \le R_0 \right\} \tag{7}$$

for all $R_0 > 0$. (One can check that the domain over which the infimum is taken, namely, the set of nonnegative measurable functions, is convex; that the objective function $\int D(R_x, \mu_x)\, d\mu(x)$ is convex on this domain; that $\int R_x\, d\mu(x)$ is a convex mapping of $\{R_x\}$ into a normed space, namely, the real line; and that there exists a rate function, namely, $R_x \equiv 0$, such that $\int R_x\, d\mu(x) < R_0$.) On the other hand, we know from (6) that $L(\lambda, \mu)$ can be interpreted as a Lagrangian for the minimization of $D(R, \mu)$ subject to a constraint on R. Hence by the Lagrange duality theorem again,

$$\begin{aligned} \max_{\lambda \ge 0} [L(\lambda, \mu) - \lambda R_0] &= \inf_R \{D(R, \mu): R \le R_0\} \\ &= D(R_0, \mu), \end{aligned} \tag{8}$$

for all $R_0 > 0$. Combining (7) and (8), we have under the conditions of Theorems 1 and 2 (and assuming $D(R, \mu_x)$ is a measurable function of R and x) the following two corollaries.

Corollary 1: $D(R, \mu) = \inf_{\{R_x\}} \{\int D(R_x, \mu_x)\, d\mu(x)$: $\int R_x\, d\mu(x) \le R\}$, $\forall R \ge 0$.

Corollary 2: $\delta^{\text{vr}}(R, \mu) = \inf_{\{R_x\}} \{\int \delta^{\text{vr}}(R_x, \mu_x)\, d\mu(x)$: $\int R_x\, d\mu(x) \le R\}$, $\forall R \ge 0$.

The special case $R = 0$ is separately and trivially proved. Corollary 1 for finite alphabets is the result by Shields, Neuhoff, Davisson, and Ledrappier [1]. Here it holds for Polish spaces. Furthermore, from our approach it can be seen that for $\lambda > 0$, each optimal R_x minimizes $D(R, \mu_x) + \lambda R$. That is, $(R_x, D(R_x, \mu_x))$ lies on the line supporting $D(R, \mu_x)$ at slope $-\lambda$. Thus, roughly speaking, the distortion-rate function of a stationary nonergodic source equals the expected value of the corresponding distortion-rate functions of its subsources, where the expectations are taken over points of equal slope.

Theorem 3 proves that variable-rate quantizers can achieve arbitrarily closely the distortion-rate bound of any stationary nonergodic source with a Polish alphabet. In contrast, fixed-rate quantizers cannot achieve arbitrarily closely the distortion-rate bound of a nonergodic source, unless almost all the ergodic modes of the source have the same distortion-rate function. Indeed, Gray and Davisson [10], [4] have shown that

$$\delta^{\text{fr}}(R, \mu) = \int \delta^{\text{fr}}(R, \mu_x)\, d\mu(x),$$

where δ^{fr} is the operational distortion-rate function for fixed-rate quantizers. This implies that

$$\begin{aligned}\delta^{\text{fr}}(R, \mu) &= \int D(R, \mu_x)\, d\mu(x) \\ &\ge \inf_{\{R_x\}} \left\{ \int D(R_x, \mu_x)\, d\mu(x) : \int R_x\, d\mu(x) \le R \right\} \\ &= D(R, \mu),\end{aligned}$$

with equality for all R only if $D(R, \mu_x) = D(R, \mu)$ for almost all x. Thus for general nonergodic sources, variable-rate quantizers are strictly more powerful than fixed-rate quantizers, even in the limit of large blocklength.

IV. Distortion-Rate Proofs

Several of the results of this paper arise from the following theorem. The theorem describes a set of conditions under which functional ergodic decomposition holds.

Theorem 4 [3, Theorem 8.9.1]: Let $(A^\infty, \mathcal{B}^\infty, \mu, T)$ be a dynamical system with a standard measurable space $(A^\infty, \mathcal{B}^\infty)$ and stationary measure μ. Let $\{\mu_x : x \in A^\infty\}$ denote the ergodic decomposition. Let $F: \mathcal{P}_s \to \Re^+$ be a nonnegative functional defined for all stationary measures and satisfying the following properties:

1) $F(\mu_x)$ is μ-integrable;
2) F is affine;
3) F is upper semicontinuous; that is, if μ_n is a sequence of stationary measures converging to a stationary measure μ in the sense that $\mu_n(G) \to \mu(G)$ for all G in a standard generating field for $\mathcal{B}^\infty$, then $F(\mu) \ge \limsup_n F(\mu_n)$.

Then

$$F(\mu) = \int F(\mu_x)\, d\mu(x).$$

We shall prove Theorem 1 using Theorem 4, with $F(\mu) = l(\lambda, \mu)$. Towards that end, we first explore some basic properties of the weighted operational distortion-rate function $l(\lambda, \mu)$ in the following lemmas. Lemmas 1–5 work towards showing that $l(\lambda, \mu)$ is affine; Lemmas 6 and 7 work towards showing that $l(\lambda, \mu)$ is upper semicontinuous.

The first lemma establishes the equivalence between two definitions of the weighted operational distortion-rate function: the first in terms of $\delta^{\text{vr}}(R, \mu)$ as in (5), and the second in terms of an Nth-order weighted operational distortion-rate function, analogous to (1) and (2).

Lemma 1: Let $l(\lambda, \mu) = \inf_R [\delta^{\text{vr}}(R, \mu) + \lambda R]$, as in (5). Then

$$l(\lambda, \mu) = \inf_N l_N(\lambda, \mu), \tag{9}$$

where

$$l_N(\lambda, \mu) = \frac{1}{N} \inf_Q E_\mu[\, \rho(X^N, Q(X^N)) + \lambda |Q(X^N)|]. \tag{10}$$

Proof: First we show that $l(\lambda, \mu) \ge \inf_N l_N(\lambda, \mu)$. Recall that $l(\lambda, \mu) = \inf_R [\delta^{\text{vr}}(R, \mu) + \lambda R]$. Given any $\epsilon > 0$, choose R such that $l(\lambda, \mu) \ge \delta^{\text{vr}}(R, \mu) + \lambda R - \epsilon$. Further, by (1) and (2), choose N and Q with expected rate $(1/N)E_\mu|Q(X^N)| \le R$ such that $\delta^{\text{vr}}(R, \mu) \ge (1/N)E_\mu \rho(X^N, Q(X^N)) - \epsilon$. Then $l(\lambda, \mu) \ge 1/N [E_\mu \rho(X^N, Q(X^N)) + \lambda E_\mu|Q(X^N)|] - 2\epsilon$. Since ϵ, N, and Q are arbitrary, the result follows. Next we show that $l(\lambda, \mu) \le \inf_N l_N(\lambda, \mu)$. Given any $\epsilon > 0$, choose N and Q such that $1/N[E_\mu \rho(X^N, Q(X^N)) + \lambda E_\mu|Q(X^N)|] \le \inf_N l_N(\lambda, \mu) + \epsilon$. Let $R = (1/N)E_\mu|Q(X^N)|$. Then $\delta^{\text{vr}}(R, \mu) \le (1/N)E_\mu \rho(X^N, Q(X^N))$, and hence $l(\lambda, \mu) \le \delta^{\text{vr}}(R, \mu) + \lambda R \le \inf_N l_N(\lambda, \mu) + \epsilon$, ϵ arbitrary. □

The next lemma shows that the infimum defining the Nth-order weighted operational distortion-rate function can be restricted to quantizers whose encoders satisfy a (biased) nearest-neighbor condition.

Lemma 2: Let $\mathcal{C} \in \mathcal{A}(N)$, where $\mathcal{A}(N)$ is the collection of finite or countable codebooks of codewords y^N with description lengths $|y^N|$ satisfying the Kraft inequality. The infimum

$$d_\lambda(x^N, \mathcal{C}) = \frac{1}{N} \inf_{y^N \in \mathcal{C}} [\, \rho(x^N, y^N) + \lambda |y^N|] \tag{11}$$

is achieved for every $x^N \in A^N$ and $\lambda > 0$. Furthermore, the infimum (10) may be restricted to those quantizers Q

for which there exists a codebook $\mathscr{C} \in \mathscr{A}(N)$ such that, for all $x^N \in A^N$,

$$Q(x^N) = \arg \min_{y^N \in \mathscr{C}} [\rho(x^N, y^N) + \lambda |y^N|]. \quad (12)$$

Thus

$$l_N(\lambda, \mu) = \inf_{\mathscr{C} \in \mathscr{A}(N)} d_\lambda(\mathscr{C}, \mu), \quad (13)$$

where

$$d_\lambda(\mathscr{C}, \mu) = E_\mu d_\lambda(X^N, \mathscr{C}). \quad (14)$$

Proof: First we establish that the infimum (11) is achieved. Arbitrarily pick $y_0^N \in \mathscr{C}$ (or more precisely, pick $y_0^N \in \{y^N: (y^N, |y^N|) \in \mathscr{C}\}$). By the Kraft inequality, there can be only a finite number of codewords $y^N \in \mathscr{C}$ with $\rho(x^N, y^N) + \lambda |y^N| \le \rho(x^N, y_0^N) + \lambda |y_0^N|$. Hence the infimum is achieved and (12) can be defined. Now for any variable-rate block quantizer Q with codewords y^N and description lengths $|y^N|$, let Q' be a variable-rate block quantizer with the same codewords and description lengths, but satisfying (12). Then

$$\rho(x^N, Q(x^N)) + \lambda |Q(x^N)| \ge \rho(x^N, Q'(x^N)) + \lambda |Q'(x^N)|,$$

so that

$$E_\mu[\rho(X^N, Q(X^N)) + \lambda |Q(X^N)|] \ge E_\mu[\rho(X^N, Q'(X^N)) + \lambda |Q'(X^N)|].$$

Hence

$$\begin{aligned} &\frac{1}{N} \inf_Q E_\mu[\rho(X^N, Q(X^N)) + \lambda |Q(X^N)|] \\ &\quad = \frac{1}{N} \inf_{Q'} E_\mu[\rho(X^N, Q'(X^N)) + \lambda |Q'(X^N)|] \\ &\quad = \frac{1}{N} \inf_{\mathscr{C} \in \mathscr{A}(N)} E_\mu \inf_{y^N \in \mathscr{C}} [\rho(X^N, y^N) + \lambda |y^N|] \\ &\quad = \inf_{\mathscr{C} \in \mathscr{A}(N)} E_\mu d_\lambda(X^N, \mathscr{C}). \end{aligned}$$

□

The next lemma shows that the Nth-order weighted operational distortion-rate function is concave, but not too concave. The lemma parallels [4, Lemma 11.2.2] for $\delta_N^{\mathrm{fr}}(R, \mu)$.

Lemma 3: For any two sources μ_1 and μ_2, any $\alpha \in [0, 1]$, any $\lambda > 0$, and any N,

$$l_N(\lambda, \alpha\mu_1 + (1-\alpha)\mu_2) \ge \alpha l_N(\lambda, \mu_1) + (1-\alpha) l_N(\lambda, \mu_2)$$

$$l_N(\lambda, \alpha\mu_1 + (1-\alpha)\mu_2) \le \alpha l_N(\lambda, \mu_1) + (1-\alpha) l_N(\lambda, \mu_2) + \frac{\lambda}{N}.$$

Proof: By (14) and the linearity of expectation,

$$\begin{aligned} &d_\lambda(\mathscr{C}, \alpha\mu_1 + (1-\alpha)\mu_2) \\ &\quad = \alpha E_{\mu_1} d_\lambda(X^N, \mathscr{C}) + (1-\alpha) E_{\mu_2} d_\lambda(X^N, \mathscr{C}) \\ &\quad = \alpha d_\lambda(\mathscr{C}, \mu_1) + (1-\alpha) d_\lambda(\mathscr{C}, \mu_2), \end{aligned}$$

for any $\mathscr{C} \in \mathscr{A}(N)$. Hence by (13),

$$\begin{aligned} &l_N(\lambda, \alpha\mu_1 + (1-\alpha)\mu_2) \\ &\quad = \inf_{\mathscr{C} \in \mathscr{A}(N)} [\alpha d_\lambda(\mathscr{C}, \mu_1) + (1-\alpha) d_\lambda(\mathscr{C}, \mu_2)] \\ &\quad \ge \alpha l_N(\lambda, \mu_1) + (1-\alpha) l_N(\lambda, \mu_2). \end{aligned}$$

On the other hand, given $\epsilon > 0$, choose $\mathscr{C}_1, \mathscr{C}_2 \in \mathscr{A}(N)$ such that $d_\lambda(\mathscr{C}_1, \mu_1) \le l_N(\lambda, \mu_1) + \epsilon$ and $d_\lambda(\mathscr{C}_2, \mu_2) \le l_N(\lambda, \mu_2) + \epsilon$. Let $\mathscr{C} = \mathscr{C}_1 \cup \mathscr{C}_2$, with the description length of each codeword increased by one bit so as to satisfy the Kraft inequality. Then

$$\begin{aligned} &l_N(\lambda, \alpha\mu_1 + (1-\alpha)\mu_2) \\ &\quad \le d_\lambda(\mathscr{C}, \alpha\mu_1 + (1-\alpha)\mu_2) \\ &\quad = \alpha d_\lambda(\mathscr{C}, \mu_1) + (1-\alpha) d_\lambda(\mathscr{C}, \mu_2), \end{aligned}$$

and for $i = 1, 2$,

$$\begin{aligned} d_\lambda(\mathscr{C}, \mu_i) &= \frac{1}{N} E_{\mu_i} \inf_{y^N \in \mathscr{C}_1 \cup \mathscr{C}_2} [\rho(X^N, y^N) + \lambda(|y^N| + 1)] \\ &\le d_\lambda(\mathscr{C}_i, \mu_i) + \frac{\lambda}{N}, \end{aligned}$$

by Lemma 2. Thus

$$\begin{aligned} &l_N(\lambda, \alpha\mu_1 + (1-\alpha)\mu_2) \\ &\quad \le \alpha d_\lambda(\mathscr{C}_1, \mu_1) + (1-\alpha) d_\lambda(\mathscr{C}_2, \mu_2) + \frac{\lambda}{N} \\ &\quad \le \alpha l_N(\lambda, \mu_1) + (1-\alpha) l_N(\lambda, \mu_2) + \frac{\lambda}{N} + \epsilon, \end{aligned}$$

ϵ arbitrary. □

The next lemma establishes the equivalence between the infimum and the limit of Nth-order weighted operational distortion-rate functions. The lemma parallels [4, Lemma 11.2.3] for $\delta^{\mathrm{fr}}(R, \mu)$.

Lemma 4: If μ is a stationary source, then

$$l(\lambda, \mu) = \lim_{N \to \infty} l_N(\lambda, \mu).$$

Proof: Recall that if a nonnegative sequence $\{a_N: N = 1, 2, \cdots\}$ is *subadditive*, i.e., if $a_{N+k} \le a_N + a_k$, then $\lim_{N \to \infty} (a_N/N) = \inf_{N \ge 1} (a_N/N)$ [3, Lemma 7.5.1]. We simply show that the nonnegative sequence $a_N = N l_N(\lambda, \mu)$ is subadditive. Fix N, k, and $\epsilon > 0$. By Lemma 2, choose variable-rate block quantizers Q_N and Q_k with blocklengths N and k whose codebooks $\mathscr{C}_N \in \mathscr{A}(N)$ and $\mathscr{C}_k \in \mathscr{A}(k)$ satisfy $d_\lambda(\mathscr{C}_N, \mu) \le l_N(\lambda, \mu) + \epsilon$ and $d_\lambda(\mathscr{C}_k, \mu) \le l_k(\lambda, \mu) + \epsilon$. Let $Q = Q_N \times Q_k$ be the blocklength-$(N + k)$ product quantizer with codebook $\mathscr{C} \in \mathscr{A}(N + k)$ that quantizes the components of $x^{N+k} =$

(x^N, x^k) separately, i.e., $Q(x^{N+k}) = (Q_N(x^N), Q_k(x^k))$ and $|Q(x^{N+k})| = |Q_N(x^N)| + |Q_k(x^k)|$. Then

$$\begin{aligned}(N+k)d_\lambda(\mathscr{C}, \mu) &= E_\mu\big[\rho\big(X^N, Q_N(X^N)\big) + \rho\big(X^k, Q_k(X^k)\big)\big] \\ &\quad + \lambda E_\mu\big[|Q_N(X^N)| + |Q_k(X^k)|\big] \\ &= Nd_\lambda(\mathscr{C}_N, \mu) + kd_\lambda(\mathscr{C}_k, \mu) \\ &\le Nl_N(\lambda, \mu) + kl_k(\lambda, \mu) + (N+k)\epsilon;\end{aligned}$$

ϵ arbitrary and $l_{N+k}(\lambda, \mu) \le d_\lambda(\mathscr{C}, \mu)$ imply

$$(N+k)l_{N+k}(\lambda, \mu) \le Nl_N(\lambda, \mu) + kl_k(\lambda, \mu),$$

i.e., $Nl_N(\lambda, \mu)$ is a subadditive sequence. □

It now follows that the weighted operational distortion-rate function $l(\lambda, \mu)$ is affine.

Lemma 5: $l(\lambda, \mu)$ is an affine function of $\mu \in \mathscr{P}_s$, the space of stationary sources.

Proof: Combine Lemmas 3 and 4. □

We will also need to show that $l(\lambda, \mu)$ is upper semicontinuous. This property is addressed in the following two lemmas.

Lemma 6 [3, Lemma 7.5.1]: Given a Polish measurable space $(A^\infty, \mathscr{B}^\infty)$ with countable generating field $\mathscr{G} = \{G_i: i = 1, 2, \cdots\}$, define a corresponding metric $d_\mathscr{G}$ on the space $\mathscr{P}(A^\infty, \mathscr{B}^\infty)$ of measures on $(A^\infty, \mathscr{B}^\infty)$ as $d_\mathscr{G}(\mu_1, \mu_2) = \sum_{i=1}^\infty 2^{-i}|\mu_1(G_i) - \mu_2(G_i)|$ for any $\mu_1, \mu_2 \in \mathscr{P}(A^\infty, \mathscr{B}^\infty)$. Let f be a nonnegative continuous function. Then $d_\mathscr{G}(\mu_n, \mu) \to 0$ implies that

$$\limsup_n E_{\mu_n} f \le E_\mu f.$$

Lemma 7: If $\rho(x_1, y_1)$ is a nonnegative continuous function of x_1 for each y_1, then $d_\lambda(x^N, \mathscr{C}) = \inf_{y^N \in \mathscr{C}} \rho(x^N, y^N) + \lambda|y^N|$ is a nonnegative continuous function of x^N for each $\mathscr{C} \in \mathscr{A}(N)$.

Proof: We need to show that if $x_n^N \to x^N$, then $d_\lambda(x_n^N, \mathscr{C}) \to d_\lambda(x^N, \mathscr{C})$. We show that there is an open neighborhood of x^N such that the infimum defining $d_\lambda(x^N, \mathscr{C})$ can be taken over a finite set. Hence in this neighborhood the infimum can be replaced by a minimum, and the needed convergence will follow. Towards this end, arbitrarily pick $y_0^N \in \mathscr{C}$, and for any $\epsilon > 0$, choose an open neighborhood around x^N such that $\rho(\tilde{x}^N, y_0^N) \le \rho(x^N, y_0^N) + \epsilon$ for all $\tilde{x}^N$ in the neighborhood. Thus in this neighborhood, $d_\lambda(\tilde{x}^N, \mathscr{C}) \le \rho(\tilde{x}^N, y_0^N) + \lambda|y_0^N|$ is bounded by a constant $K < \infty$. Since the description lengths $|y^N|$ satisfy the Kraft inequality, there can be only a finite number of codewords $y^N \in \mathscr{C}$ such that $\lambda|y^N| \le K$. Indeed, y_0^N is in this set, and the infimum can be replaced by a minimum over this set for all $\tilde{x}^N$ in the neighborhood. □

We now turn our attention to the proof of Theorem 1. The proof parallels the proof of [4, Theorem 11.3.1] of the ergodic decomposition of $\delta^{\text{fr}}(R, \mu)$.

Proof of Theorem 1: Show that $l(\lambda, \mu)$ satisfies the conditions of Theorem 4:

1) *Integrable.* $l(\lambda, \mu_x) \le l_N(\lambda, \mu_x) \le E_{\mu_x}\rho(X_1, y_1^*)$ implies $l(\lambda, \mu_x)$ is integrable, since $\rho(X_1, y_1^*)$ is.
2) *Affine.* By Lemma 5.
3) *Upper semicontinuous.* Since A and hence A^∞ is Polish, choose distance $d_\mathscr{G}(\mu_1, \mu_2) = \sum_{i=1}^\infty 2^{-i}|\mu_1(G_i) - \mu_2(G_i)|$ for any $\mu_1, \mu_2 \in \mathscr{P}(A^\infty, \mathscr{B}^\infty)$, where $\mathscr{G} = \{G_i: i = 1, 2, \cdots\}$ is a countable standard generating field of $\mathscr{B}^\infty$. Pick N large enough such that $l_N(\lambda, \mu) \le l(\lambda, \mu) + \epsilon$ and such that there exists a codebook $\mathscr{C} \in \mathscr{A}(N)$ for which $d_\lambda(\mathscr{C}, \mu) \le l_N(\lambda, \mu) + \epsilon$. Then, given some $\{\mu_n\}$ for which $d_\mathscr{G}(\mu_n, \mu) \to 0$ as $n \to \infty$,

$$\begin{aligned}\limsup_n l(\lambda, \mu_n) &\le \limsup_n l_N(\lambda, \mu_n) \\ &\le \limsup_n d_\lambda(\mathscr{C}, \mu_n) \\ &\le E_\mu d_\lambda(X_N, \mathscr{C}) \text{ by Lemmas 6 and 7} \\ &\le l_N(\lambda, \mu) + \epsilon \\ &\le l(\lambda, \mu) + 2\epsilon,\end{aligned}$$

ϵ arbitrary, which implies $l(\lambda, \mu)$ is upper semicontinuous for fixed λ. □

We finally turn our attention to the ergodic decomposition of $L(\lambda, \mu)$. We will need the following lemma.

Lemma 8: Let $L(\lambda, \mu) = \inf_R [D(R, \mu) + \lambda R]$, as in (6). Then

$$L(\lambda, \mu) = \inf_N L_N(\lambda, \mu),$$

where

$$L_N(\lambda, \mu) = \frac{1}{N} \inf_\nu \big[E_{\mu\nu}\rho(X^N, Y^N) + \lambda I_{\mu\nu}(X^N; Y^N)\big].$$

Proof: Similar to the proof of Lemma 1. □

Proof of Theorem 2: By the converse to the source coding theorem [6, Theorem 7.2.5], $D(R, \mu) \le \delta^{\text{vr}}(R, \mu)$, so

$$\begin{aligned}L(\lambda, \mu) \le l(\lambda, \mu) &= \int l(\lambda, \mu_x)\, d\mu(x) \\ &= \int L(\lambda, \mu_x)\, d\mu(x).\end{aligned}$$

Thus we need only prove that $L(\lambda, \mu) \ge \int L(\lambda, \mu_x)\, d\mu(x)$. By Lemma 8, choose a test channel ν such that $(1/N)[E_{\mu\nu}\rho(X^N, Y^N) + \lambda I_{\mu\nu}(X^N; Y^N)] \le L_N(\lambda, \mu) + \epsilon$. Then

$$\begin{aligned}&\int L_N(\lambda, \mu_x)\, d\mu(x) \\ &\quad \le \frac{1}{N}\int \big[E_{\mu_x\nu}\rho(X^N, Y^N) + \lambda I_{\mu_x\nu}(X^N; Y^N)\big]\, d\mu(x) \\ &\quad \le \frac{1}{N}\big[E_{\mu\nu}\rho(X^N, Y^N) + \lambda I_{\mu\nu}(X^N; Y^N)\big] \\ &\quad \le L_N(\lambda, \mu) + \epsilon,\end{aligned}$$

where the second inequality follows from the concavity of $I_{\mu_x\nu}(X^N; Y^N)$ in μ_x [5]. Thus

$$\inf_N \int L_N(\lambda, \mu_x)\, d\mu(x) \le L(\lambda, \mu) + \epsilon,$$

and hence, using Fatou's lemma,

$$\begin{aligned}\int L(\lambda, \mu_x)\, d\mu(x) &= \int \inf_N L_N(\lambda, \mu_x)\, d\mu(x) \\ &\le \inf_N \int L_N(\lambda, \mu_x)\, d\mu(x) \\ &\le L(\lambda, \mu) + \epsilon,\end{aligned}$$

ϵ arbitrary. □

Acknowledgment

We would like to thank the anonymous referees for their constructive (and in one case, face-saving) suggestions.

References

[1] P. C. Shields, D. L. Neuhoff, L. D. Davisson, and F. Ledrappier, "The distortion-rate function for nonergodic sources," *Ann. Probab*, vol. 6 no. 1, pp. 138–143, 1978.

[2] A. Leon-Garcia, L. D. Davisson, and D. L. Neuhoff, "New results on coding of stationary nonergodic sources," *IEEE Trans. Inform. Theory*, vol. IT-25 no. 2, pp 137–144, Mar. 1979.

[3] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag, 1988.

[4] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[5] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[6] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[7] J. C. Kieffer, "On the optimum average distortion attainable by a fixed-rate coding of a nonergodic source," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 190–193, Mar. 1975.

[8] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

[9] T. H. Chung, "Minimax learning in iterated games via distributional majorization," Ph.D. Dissertation, Stanford University, 1994.

[10] R. M. Gray and L. D., Davisson, "Source coding theorems without the ergodic assumption," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 502–526, July 1974.